ESTUDAM Bilişim Derg., 2024; 5(1); 29-36

J of ESTUDAM Information, 2024; 5(1): 29-36.

DOI: 10.53608/estudambilisim. 1293834

(Geliş Tarihi / Received Date: 10.05.2023, Kabul Tarihi/ Accepted Date: 22.01.2024)

(**Araştırma Makalesi**)

Hesaplamalı İlaç Keşfi ve Makine Öğrenme Algoritmaları

Amin Hashemian[1], Giyasettin OZCAN[2]

[1] Bursa Uludag Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 16059, Bursa, ORCID No: http://orcid.org/0009-0007-9591-4217

[*2] Bursa Uludag Üniversitesi, Mühendislik Fakültesi, Bilgisayar Bölümü, 16059, Bursa, ORCID No : https://orcid.org/0000-0002-1166-5919

Anahtar Kelimeler:

Hesaplamalı ilaç keşfi,
Moleküler modelleme,
Makine öğrenmesi,
Yapı-tabanlı ilaç tasarımı,
Derin öğrenme,
Moleküler
dinamik simülasyonlar

Özet: Hesaplamalı ilaç keşfi, geleneksel laboratuvar yöntemleri ve deneysel çalışmaların birlikte analiz edilmesini amaçlamaktadır ve ilaç keşif sürecinde önemli bir rol oynamaktadır. Bu çalışmada, hesaplamalı yöntemlerin ilaç keşfi alanında nasıl kullanıldığına odaklanılmaktadır. İlk olarak, moleküler modelleme ve simülasyon tekniklerinin, ilaç adayı bileşiklerin tasarımı ve özelliklerinin anlaşılması için nasıl kullanıldığı anlatılmaktadır. Moleküler dinamik simülasyonlar ve yapı-tabanlı ilaç tasarımı gibi yöntemler, potansiyel ilaç moleküllerinin etkileşim mekanizmalarını ve hedef proteinlerle ilişkilerini incelemektedir.

Makalenin ikinci bölümünde, sanal tarama yöntemleri ele alınmaktadır. Sanal tarama yöntemleri, hedef proteinin yapısını kullanarak, potansiyel bağlanma bölgelerini ve etkileşim alanlarını tahmin ederek, ilaç adayı moleküllerin seçiminde ve optimize edilmesinde önemli bir rol oynamaktadır. Son olarak, makalenin üçüncü bölümünde, makine öğrenmesi ve yapay zeka tekniklerinin ilaç keşfi alanında nasıl kullanıldığı tartışılmaktadır. Bu amaçla moleküler tasarım sürecinde yeni moleküllerin üretilmesinde ve ilaçların etkileşim mekanizmalarının anlaşılması incelenmiştir ve ilaç keşfi konusunda tahmin yapan bir uygulama sunulmuştur. Bu amaçla TP53 gen varyasyonlarının ilaç etkileşimleri analiz edilmiştir.

(**Research Article**)

Computational Drug Discovery and Machine Learning Algorithms

Keywords:

Computational drug discovery,
Molecular modeling,
Machine learning,
Structure-based drug design,
Deep learning,
Molecular dynamics
simulations

Abstract: Computational Drug Discovery plays a significant role in the drug discovery process when used in conjunction with traditional laboratory methods and experimental studies. This study focuses on how computational methods are employed in the field of drug discovery. Firstly, it describes how molecular modeling and simulation techniques are utilized for the design and understanding of the properties of drug candidate compounds. Methods such as molecular dynamics simulations and structure-based drug design are commonly used to investigate the interaction mechanisms of potential drug molecules and their relationships with target proteins.

In the second section, virtual screening methods are addressed. Virtual screening methods play a crucial role in the selection and optimization of drug candidate molecules by predicting potential binding sites and interaction areas based on the structure of the target protein. Finally, machine learning and artificial intelligence techniques are discussed in the field of drug discovery. For this purpose, the generation of new molecules in the molecular design process and understanding the interaction mechanisms of drugs are examined. In this study, an application that predicts drug discovery is developed and presented. For this purpose, drug interactions of TP53 gene variations were analyzed.

* Sorumlu yazar/Corresponding author: gozcan@uludag.edu.tr

1. GİRİŞ

İlaç keşfi, insan sağlığının iyileştirilmesi amacıyla yürütülen bir süreçtir. Geleneksel ilaç keşif süreci genellikle zaman alıcı, maliyetli ve deneysel çalışmaları içeren bir süreçtir. Bu nedenle, bilim insanları ve araştırmacılar, ilaç keşfi sürecini hızlandırmak ve daha verimli hale getirmek için yeni yöntemler arayışı içindedir.
Hesaplamalı ilaç keşfi, son yıllarda ilaç keşfi alanında büyük bir ilgi çekmiştir. Hesaplamalı yöntemler, ilaç adayı bileşiklerin tasarımı, moleküler özelliklerinin belirlenmesi, etkileşim mekanizmalarının anlaşılması ve ilaç adayı seçimi gibi alanlarda değerli bir araç olmuştur. Moleküler modelleme, simülasyon teknikleri, sanal tarama ve makine öğrenmesi gibi hesaplamalı yaklaşımlar, ilaç keşfinde hızlı, ekonomik ve verimli bir şekilde kullanılabilmektedir [1].

İlaç keşfi, bütünsel tedaviye odaklanan geleneksel bir yaklaşımdır. Geçen yüzyılda, dünyadaki tıp toplulukları tedavi ve iyileşme için geleneksel bir yaklaşım kullanmaya başlamıştır. Günümüzde, Makine Öğrenimi (ML) ve Derin Öğrenme (DL), ilaç keşfi için çekici yaklaşımlar haline gelmiştir. Kişiselleştirilmiş ilaç tedavisinin ana konsepti, bireysel genomik profillere dayalı ilaçlar önermektir. Daha önce ilaç tedavilerinin çoğu, hastalığın anatomik kökenine dayanıyordu, ancak daha sonra moleküler analizler şunu netleştirmiştir: Hastanın genomik karakterizasyonu, ilaç tedavisinin tasarlanmasında önemli bir rol oynamaktadır. Çeşitli yapay öğrenme yaklaşımları ile ilaç kombinasyonunu tahmin etmek için literatürde önerilmiştir [2]. Bu yaklaşımların çoğu, ilaç duyarlılığını tahmin etmek için genetik ölçümlere dayanır ve hücre hatlarındaki ilaç yapı benzerliğini ve gen ifadesi benzerliğini kullanmaktadır.

Mevcut araştırmada AI ilaç keşfi hakkındaki incelemeler üç kategoriye ayırarak kısaca tartışılmıştır:

1) "Genel ilaç keşif incelemesi",
2) "AI çağında ilaç keşfi"
3) " Veri, Temsil Ve Kıyaslama Platformları".

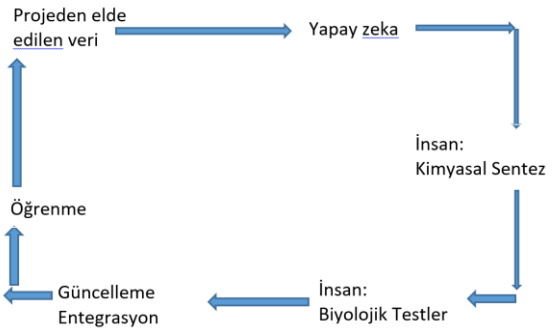
Ayrıca, geleneksel laboratuvar yöntemleriyle birlikte hesaplamalı yöntemlerin nasıl birleştirildiği ve bu birleşimin ilaç keşfindeki potansiyeli de ele alınmaktadır. Özellikle, moleküler modelleme ve simülasyon teknikleri, sanal tarama yöntemleri ve makine öğrenmesi algoritmalarının ilaç keşfi sürecindeki önemi vurgulanmaktadır.

2.1. Hesaplamalı İlaç Keşfi

İlaç keşfi, bir hastalık için herhangi bir ilacın bilinmediği durumlarda mevcut ilaçların sınırlı etkinliğini ve şiddetli toksisitesini araştırır [2]. En erken aşamada, bir hedefin (örn., bir enzim, bir reseptör, bir iyon kanalı) aktivasyonunun veya inhibisyonunun hastalık için hedef tanımlama ve hedef doğrulamayı içeren terapötik etkilerle sonuçlandığına dair temel bir hipotez geliştirilmelidir [2]. Seçilen hedef için olası ilaç adaylarını bulmak için yoğun testler yapılmalıdır. İlaç adayları daha sonra klinik öncesi çalışmalara ve klinik araştırmalara girerler. Başarılı olursa, ilaç adayı hastalığı tedavi etmek için tıbbi bir ürün olarak piyasaya sürülebilir. Küçük moleküllü ilaç keşfini hızlandırmak için, 1980 'lerden bu yana yüksek verimli tarama (HTS) önerilmiştir [3]. HTS 'nin öne çıkan bir sonucu, PubChem ve ZINC gibi kimyasal veri tabanlarına katkıdır. Sanal tarama (VS) olarak da bilinen deneylerde test edilecek potansiyel aktif moleküller için kimyasal kütüphaneleri aramak için çeşitli hesaplama teknikleri geliştirilmiştir [4]. Başka bir deyişle, VS, aktif molekülleri tanımlama olasılığını artırmak için, hedef (yapıya dayalı VS) veya bilinen aktif (ligand tabanlı VS), hakkındaki bilgilere dayanan hesaplama yaklaşımlarını kullanarak aktif molekülleri tanımlamaktır. Bilgisayar destekli ilaç keşif (CADD) araçları, bu süreçleri otomatikleştirmek ve hızlandırmak ve araştırma ve geliştirme maliyetini azaltmak için kullanılabilir. Biyoinformatik araştırmaları, biyolojik yapılar, moleküler veri tabanları ilaç keşfi ve geliştirme hattının çeşitli aşamalarında kullanılabilecek çeşitli hesaplama araçları gibi önemli miktarda veri kaynağını kullanıma sunmuştur. İlaç keşfi, büyük sosyal ve ticari etkilere sahip önemli bir gelişmedir. Yeni bir ilacı geleneksel yollarla keşfetmek e onlarca yıl süre ve milyarlarca dolar harcama gerektirmektedir. Yapılan araştırmaların umut verici ve heyecan verici bir yönü, ilaç keşfini hızlandırmak için yapay zekanın (AI) kullanılmasıdır. Yapay zeka (AI) son on yılda ilaç keşfine büyük bir katkıda bulunmuş. Sanal tarama ve ilaç tasarımı gibi birçok ilaç keşif uygulamasında çeşitli AI teknikleri kullanılmıştır.

2.2. Yapay Zekâ Çağında İlaç Keşfi

AI, ilaç keşfinde yaygın olarak uygulanmaktadır. 2000'li yılların başından beri, Random Forest (RF) gibi makine öğrenimi modelleri VS ve QSAR için kullanılmıştır. 2012'de AlexNet [5], derin öğrenme çağının başlangıcına işaret etmiştir. 2012 Merck Kaggle yarışmasından kısa bir süre sonra, derin sinir ağları (DNN), moleküler aktiviteleri tahmin etmede standart RF modelinden daha iyi performans göstermiştir. Daha yakın zamanlarda, yapay zeka tekniklerinin bilgisayarla görme ve doğal dil işleme deki başarısı, ilaç keşfine daha fazla ışık tutmuş ve kimyada gelişen derin öğrenme alanına yol açmıştır [6]. 2019 yılında, Insilico Medicine araştırmacıları tarafından 21 gün içinde güçlü diskoidin alan reseptörü (discoidin domain receptor1) inhibitörleri keşfedilmiştir [7]. 2020 yılında, antibiyotiğe dirençli bakterilere karşı yeni bir antibiyotik adayı olan Halicin, MIT araştırmacıları tarafından tanımlanmıştır [8].
Öte yandan yapay zekanın ilaç keşfi konusundaki vaadine rağmen, yaygın olarak tartışılan sorunlar hala mevcuttur [9]. Hem ilaç keşif uygulamaları hem de yapay zekâ teknikleri hakkında net bir anlayışın gerekliliği göz önüne alındığında, çoğu araştırma yazıları ilaç keşfindeki genel yönlerden başlar ve ardından veri kaynaklarını, molekül temsillerini, model mimarilerini ve öğrenme paradigmalarını kapsayan yapay zekâ güdümlü ilaç keşfine geçer. Bu incelemenin organizasyonu Şekil 1 'de gösterilmektedir.



Şekil 1. İlaç Keşfinde Yapay zeka .

Hesaplamalı ilaç keşfinde, son yıllarda evrişimli sinir ağları derin öğrenme yöntemlerinin önemli bir rol oynadığı bir alandır. Evrişimli sinir ağları, moleküler yapıların analizi ve ilaç adayı keşfi süreçlerinde kullanılan güçlü bir araçtır. Evrişimli Sinir Ağları (CNN 'ler) genellikle görüntü işleme, özellikle de bilgisayarlı görü veya görsel veri analizi alanında veri piksellerini işlemek için kullanılır [10]. CNN 'lerde evrişim katmanları ve havuzlama (yani alt örnekleme) katmanları vardır. Bu evrişim katmanlarının ve havuzlama katmanlarının üzerinde, son tahmin için özellik haritalarını birleştirerek bir vektör temsili öğrenilir. CNN'ler, öğrenilecek parametre sayısını büyük ölçüde azaltan, böylece bellek tüketimini azaltan ve hesaplama hızını artıran filtreler arasında parametreleri paylaşır.

Evrişimli sinir ağları (CNN 'ler), derin öğrenme alanında popüler hale gelen bir yapay sinir ağı türüdür. Moleküler yapıların analizi için kullanıldığında, CNN'ler moleküllerin grafiksel temsillerini girdi olarak alır ve çeşitli katmanlar boyunca veriyi işler. Bu katmanlar, moleküler özelliklerin öğrenilmesi, temsil edilmesi ve anlaşılması için özelleştirilmiştir. Evrişimli sinir ağları, hesaplamalı ilaç keşfi için çeşitli alanlarda kullanılmaktadır. Öncelikle, moleküler aktivite tahmini ve QSAR analizi gibi ilaç etkileşimlerinin keşfi için kullanılırlar. CNN'ler, moleküler yapıların özelliklerini ve etkileşim desenlerini yakalamak için moleküldeki atomlar, bağlar ve alt yapılar arasındaki ilişkileri modelleyebilir. Bu, potansiyel ilaç adaylarının aktivitelerini tahmin etmek ve uygunluk analizleri yapmak için kullanılabilir. Evrişimli sinir ağları, ilaç keşfi sürecinde sanal tarama ve moleküler tasarım için de kullanılmaktadır. Örneğin, moleküler tarama yöntemleri ile kimyasal kütüphaneler taranırken, CNN'ler moleküllerin benzerliklerini ve özelliklerini analiz ederek ilaç adaylarının seçilmesine yardımcı olabilir. Bunun yanı sıra, moleküler tasarım sürecinde, evrişimli sinir ağları yeni ilaç adayı moleküllerin olası yapılarını üretmek ve optimize etmek için kullanılabilir.

2015 yılında Duvenaud [11], sabit kimyasal tanımlayıcılar yerine moleküler özellik tahmini için veriye dayalı temsil öğrenimini kullanan ilk çabalardan biri olan farklılaştırılabilir bir parmak izi oluşturmak için ECFP 'lerin [12] bir iyileştirmesi olan dairesel parmak izlerine CNN'leri uyguladı. Bu çalışma, öğrenme moleküler temsillerini büyük ölçüde geliştirdi. Parmak izlerine ek olarak, CNN'ler ayrıca doğrudan moleküler yapı görüntülerinden özellikleri etkili bir şekilde çıkarabilir. Örneğin, Chemception [13], HIV replikasyonunun inhibisyonunu ve solvasyonun serbest enerjisini tahmin etmek için 2D-yapısal görüntüler üzerinde eğitilmiştir. Daha sonra, Fernandez [14], girdi olarak görüntülerle birlikte toksisite sınıflandırması için bir çerçeve olan Toxic Colours 'u geliştirdi. Bunlarla birlikte, evrişimli sinir ağları, hesaplamalı ilaç keşfinde gelecekte daha da geliştirilebilir ve iyileştirilebilir. Veri setlerinin büyümesi ve çeşitliliği ile birlikte, CNN'ler daha güçlü ve hassas modeller oluşturabilir. Ayrıca, farklı veri temsilleri, özellik mühendisliği teknikleri ve transfer öğrenme gibi ileri tekniklerin kullanılması, evrişimli sinir ağlarının performansını artırabilir.

Sonuç olarak, evrişimli sinir ağları, hesaplamalı ilaç keşfi sürecinde veri analizi, aktivite tahmini, sanal tarama ve moleküler tasarım gibi alanlarda önemli bir araçtır. Bu ağların yüksek öğrenme kapasiteleri ve derin öğrenme yetenekleri, ilaç adayı keşfinde hızlı ve verimli sonuçlar elde etmeyi mümkün kılar. Ancak, bu tekniklerin kullanımı, uygun veri yönetimi, eğitim süreci ve ilaç keşfi sürecinin diğer unsurlarıyla entegrasyonunu gerektirir. Gelecekte, evrişimli sinir ağlarının daha da geliştirilmesi ve optimize edilmesi, hesaplamalı ilaç keşfi sürecinde büyük potansiyeller sunma potansiyeline sahiptir.

2.3. Veri, Temsil ve Kıyaslama Platformları

Moleküller genellikle bağları ve atomları olan Kekul´e diyagramları olarak tasvir edilir moleküllerin hızlı hesaplanmasını, sorgulanmasını ve depolanmasını sağlamak için makine tarafından okunabilen temsiller geliştirilmiştir [15].

Hesaplamalı ilaç keşfi için Veri, Temsil ve Kıyaslama Platformları, ilaç adayı keşfi ve tasarımı sürecinde önemli bir rol oynamaktadır. Bu platformlar, büyük miktarda veriyi işleme, moleküler yapıları temsil etme ve ilaç adaylarını karşılaştırma yetenekleri sunar. Bu metinde, hesaplamalı ilaç keşfi için Veri, Temsil ve Kıyaslama Platformlarına odaklanarak, önemli birçok yöntem ve araçtan bahsedilecektir.
Hesaplamalı ilaç keşfi sürecinde, genellikle büyük veri kümeleri kullanılır. Veri platformları, bu verilerin depolanması, yönetimi ve erişimi için önemli bir rol oynar. Kimyasal veri tabanları, ilaç adayı bileşiklerin yapılarını, özelliklerini ve etkileşim bilgilerini barındırır. Bunlar, yapı aktivite ilişkisi (structure-activity relationship) analizi, farmakofor tabanlı tarama ve ilaç etkileşim mekanizmalarının keşfi gibi hesaplamalı yöntemler için önemli bir veri kaynağı sağlar. Moleküler yapıların doğru ve anlamlı bir şekilde temsil edilmesi, hesaplamalı ilaç keşfinin temelidir. Temsil platformları, moleküler yapıların matematiksel veya grafiksel bir şekilde kodlanmasını sağlar. Kimyasal temsiller, moleküllerin atomları, bağları ve diğer yapısal özelliklerini içerir. Bu temsiller, moleküler özelliklerin analizinde, sanal tarama yöntemlerinde ve makine öğrenmesi algoritmalarında kullanılır. İlaç adayı moleküllerin karşılaştırılması ve benzerlik analizi önemli bir adımdır. Kıyaslama platformları, moleküler yapıların

benzerliklerini değerlendirmek ve ilaç adaylarını sıralamak için kullanılır.
Kimyasal benzerlik arama, moleküler tanımlayıcılar ve algoritmalar kullanarak moleküler yapıları karşılaştırır. Bu platformlar, potansiyel ilaç adaylarının hızlı bir şekilde tespit edilmesini ve seçilmesini sağlar.

3. MATERYAL VE METOT

Bu araştırma kapsamında ChEMBL biyoaktivite verilerini kullanarak bir makine öğrenmesi modeli oluşturulmuştur. İlk aşamada ChEMBL veri tabanından veri toplama ve ön işleme gerçekleştirilmiştir. Araştırma kapsamında P53 proteinin mutasyonları üzerinde makine öğrenme modeli ve web arayüzü gerçekleşmiştir.

3.1. Chembl Veritabanı

ChEMBL veri tabanı, 2 milyondan fazla bileşiğin derlenmiş biyoaktivite verilerini içeren bir veri tabanıdır. 76.000'den fazla belgeden, 1,2 milyon tahlilden derlenmiştir ve veriler 13.000 hedefi, 1.800 hücreyi ve 33.000 belirtiyi kapsar [16]. ChEMBL veri tabanından biyoaktivite verilerini alabilmemiz için ChEMBL web hizmet paketi kurulmalıdır.

3.2. Tümör Protein P53

Tümör proteini P53 olarak da bilinen p53, hücresel tümör antijeni p53 veya transformasyonla ilgili protein (TP53) olarak tanımlanmaktadır, ve insan kanserlerinde sıklıkla mutasyona uğrayan düzenleyici bir proteindir. p53 proteinleri, kanser oluşumunu önledikleri omurgalılar için çok önemlidir [17]. İnsanlarda TP53 geni, 17. kromozomun kısa kolunda bulunur. TP53 geni hasar görürse, tümör baskılanması ciddi şekilde tehlikeye girer. TP53 geninin yalnızca bir işlevsel kopyasını miras alan kişilerde, büyük olasılıkla erken yetişkinlik döneminde Li-Fraumeni sendromu olarak bilinen bir hastalık olan tümörler gelişecektir. TP53 geni, kontrolsüz hücre bölünmesi olasılığını artıran mutajenler (kimyasallar, radyasyon veya virüsler) tarafından da modifiye edilebilir. İnsan tümörlerinin yüzde 50'den fazlası, TP53 geninin bir mutasyonunu veya silinmesini içerir [18]. p53'ün kaybı, çoğunlukla bir anöploidi fenotipi ile sonuçlanan genomik istikrarsızlık yaratır. Tümörlerin tedavisi veya yayılmasının önlenmesi için p53 miktarının artırılması bir çözüm gibi görünebilir.
Bu çalışmada p53 proteininin farklı mutasyonları üzerinde araştırma yapılmıştır. Bu mutasyonlardan hangilerinin girdiği kimyasal reaksiyonda daha etkili olduğu analiz edilmiştir ve buna uygun makine öğrenmesi modelleri uygulaması geliştirilmiştir. Bu açıdan yola çıkarak bir proteinin mutasyon kümesinden en önemlilerini bulmak aslında bu hastalıkla ilgili en verimli ilacı bulmak demektir bu ilaç keşfine yol açabilir.

Yapılan uygulamanın aşamalarında ilk adımda tümör ile ilgili proteinler ChEMBL veritabanında arandı ve bu protein kümesinden p53 mutasyonları seçildi sonuçlar csv dosyası halinde kayıt edildi ve P53 protein kümesinde IC50 değerine sahip proteinler filtrelendi. Eksik veriler silindikten ve veriler üzerinde ön işleme yaptıktan sonra veri kümesinin son hali Şekil 2' ve Şekil 3't eki gibi olmuştur:

```
In [10]: selection = ['molecule_chembl_id','canonical_smiles','standard_value']
         df3 = df2_nr[selection]
         df3
```

Out[10]:

	molecule_chembl_id	canonical_smiles	standard_value
0	CHEMBL191334	COc1ccc(C2=N[C@@H](c3ccc(Cl)cc3)[C@@H](c3ccc(C...	1390.0
1	CHEMBL583413	Cc1ccc2c(c1)C1=C(C(c3ccc(Br)cc3)O2)C(c2ccc(Br)...	3000.0
2	CHEMBL584512	Cc1cc2c(cc1F)OC(c1ccc(Br)cc1)C1=C2N(C)c2ncnn2C...	11400.0
3	CHEMBL567403	CN1C2=C(C(c3ccc(Br)cc3)Oc3ccc(F)cc32)C(c2ccc(B...	44300.0
4	CHEMBL583626	CCOc1cccc2c1C1=C(C(c3ccc(Br)cc3)O2)C(c2ccc(Br)...	6500.0
5	CHEMBL567636	CN1C2=C(C(c3ccc(Cl)cc3)Oc3ccccc32)C(c2ccc(Br)c...	15000.0
6	CHEMBL584299	CN1C2=C(C(c3ccc(Cl)cc3)Oc3ccccc32)C(c2ccc(Cl)c...	9370.0
9	CHEMBL2059300	CC[C@@H](C(=O)OC(C)(C)C)N1C(=O)[C@@](C)(CC(=O)...	1000.0
11	CHEMBL2059435	CC[C@@H]([C@H](C)O)N1C(=O)[C@@](C)(CC(=O)O)C[C...	760.0
12	CHEMBL2177191	CC[C@@H](CO)N1C(=O)[C@@](C)(CC(=O)O)C[C@H](c2c...	1500.0
13	CHEMBL2177190	CCC(CC)N1C(=O)[C@@](C)(CC(=O)O)C[C@H](c2cccc(C...	5500.0
14	CHEMBL2177189	CC[C@@H](C(=O)OC(C)(C)C)N1C(=O)[C@@](C)(Cc2nnn...	1600.0

Şekil 2. Chembl Veri Tabanı P53 Proteini Veri Ön İşlemesi Gösterimi.

Biyoaktivite verileri IC50 biriminde ifade edilir. 1000 nM 'den daha düşük değerlere sahip bileşiklerin aktif olduğu kabul edilirken, 10.000 nM 'nin üzerindeki bileşiklerin aktif olmadığı kabul edilecektir. 1.000 ve 10.000 nM arasındaki bu değerlere gelince, ara olarak anılacaktır.

```
In [13]: bioactivity_threshold = []
         for i in df4.standard_value:
           if float(i) >= 10000:
             bioactivity_threshold.append("inactive")
           elif float(i) <= 1000:
             bioactivity_threshold.append("active")
           else:
             bioactivity_threshold.append("intermediate")

In [14]: bioactivity_class = pd.Series(bioactivity_threshold, name='class')
         df5 = pd.concat([df4, bioactivity_class], axis=1)
         df5
```

Out[14]:

	molecule_chembl_id	canonical_smiles	standard_value	class
0	CHEMBL191334	COc1ccc(C2=N[C@@H](c3ccc(Cl)cc3)[C@@H](c3ccc(C...	1390.0	intermediate
1	CHEMBL583413	Cc1ccc2c(c1)C1=C(C(c3ccc(Br)cc3)O2)C(c2ccc(Br)...	3000.0	intermediate
2	CHEMBL584512	Cc1cc2c(cc1F)OC(c1ccc(Br)cc1)C1=C2N(C)c2ncnn2C...	11400.0	inactive
3	CHEMBL567403	CN1C2=C(C(c3ccc(Br)cc3)Oc3ccc(F)cc32)C(c2ccc(B...	44300.0	inactive
4	CHEMBL583626	CCOc1cccc2c1C1=C(C(c3ccc(Br)cc3)O2)C(c2ccc(Br)...	6500.0	intermediate
5	CHEMBL567636	CN1C2=C(C(c3ccc(Cl)cc3)Oc3ccccc32)C(c2ccc(Br)c...	15000.0	inactive
6	CHEMBL584299	CN1C2=C(C(c3ccc(Cl)cc3)Oc3ccccc32)C(c2ccc(Cl)c...	9370.0	intermediate
7	CHEMBL2059300	CC[C@@H](C(=O)OC(C)(C)C)N1C(=O)[C@@](C)(CC(=O)...	1000.0	active

Şekil 3. Chembl Veri Tabanı P53 Proteini Veri Ön İşlemesi Gösterimi.

Smiles formatı, bir molekülün bağlanabilirliğini ve kiralitesini tanımlayabilen doğrusal bir metin formatıdır. kanonik smıles, herhangi bir belirli molekül için tek bir "kanonik" form verir. Bu çalışmada moleküllerin canonical smiles formatları üzerinde hesaplama yapılmıştır ve yapay zeka modelinde de girdi olarak moleküllerin canonical smiles formatı dikkate alınmıştır.

3.3. Lipinski Tanımlayıcılarını Hesapla

Bileşiklerin ilaca benzerliğini değerlendirmek için bir dizi pratik kural Lipinski kuralı olarak bilinmektedir [19]. Bir ilaca benzerlik, farmakokinetik profil olarak da bilinen Absorpsiyon, Dağılım, Metabolizma ve Atılım 'a (ADME) dayanır. Lipinski, Kuralı olarak bilinen formülasyonda aktif tüm FDA onaylı ilaçları analiz etmiştir.

Lipinski Kuralı aşağıdakileri ifade eder:

· Molekül ağırlığı < 500 Dalton
· Oktanol-su dağılım katsayısı (LogP) < 5
· Hidrojen bağı donörleri < 5
· Hidrojen bağı alıcıları < 10

Kısaca beş kuralı (RO5) olarak da bilinen Lipinski 'nin beş kuralı, ilaca benzerliği değerlendirmek veya belirli bir farmakolojik veya biyolojik aktiviteye sahip bir kimyasal bileşiğin kimyasal özelliklere ve fiziksel özelliklere sahip olup olmadığını belirlemek için kullanılan pratik bir kuraldır. Kural, 1997'de Christopher A. Lipinski tarafından, ağızdan uygulanan ilaçların çoğunun nispeten küçük ve orta derecede lipofilik moleküller olduğu gözlemine dayanarak formüle edildi. Kural, emilim, dağılım, metabolizma ve atılım dahil olmak üzere bir ilacın insan vücudundaki farmakokinetiği (vücutta ilaç hareketi) için önemli olan moleküler özellikleri tanımlar. Ancak kural, bir bileşiğin farmakolojik olarak aktif olup olmadığını öngörmez [19].
Yapılan çalışmada verilerin canonical smiles formatı üzerinde hesaplama yapılmış ve sonuçlar Şekil 4'te gösterildiği üzere matrisler halinde kaydedilmiş ve bu veriler bir sonraki aşamada Ml modeli oluşturmak için kullanılmıştır.

X data matrix

```
In [0]: df3_X = pd.read_csv('descriptors_output.csv')
In [16]: df3_X
```

Out[16]:

	Name	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4
0	CHEMBL336398	1	1	1	0	0
1	CHEMBL133897	1	1	1	0	0
2	CHEMBL130628	1	1	1	0	0
3	CHEMBL131588	1	1	0	0	0
4	CHEMBL130478	1	1	0	0	0

Şekil 4. Sonuç Matrislerinin Gösterimi

Son bölümde, random forest algoritması kullanılarak p53 proteinin bir regresyon modelini inşa edilmiştir. p53 veri seti 881 giriş özelliği ve 1 çıkış değişkeni (pIC 50 değerleri) içerir. Verilerinin daha düzgün dağılmasına izin vermek için, IC50'yi temelde -log10(IC50) olan negatif logaritmik ölçeğine çevrilmiştir. Bu değer pIC 50 değerlerine eşittir.

QSAR modelleri, bir dizi molekülün fizikokimyasal özellikleri (moleküler tanımlayıcılar olarak adlandırılır) ile farmakolojik veya biyolojik aktiviteyi ilişkilendiren matematiksel modellerdir. Çalışmada jupyter notebook ta yapılan sonuçları kullanarak geliştirilmiş model web uygulamasında kullanabilmek için pikle objesi şeklinde kaydedilmiştir. Önceki bölümlerde yapılan geliştirmelerin son amacı bir QSAR modeli oluşturmaktır.

3.4. Streamlit Web Uygulaması

Bu bölümde geliştirilmiş yapay zeka modeli web uygulaması içinde kullanılabilmesi için app.py dosyasında Moleküler tanımlayıcı hesaplama, Dosya indirme, Model oluşturma gibi fonksiyonlar geliştirilmiştir. Molekül tanımlayıcı, dosya indirme ve model oluşturma fonksiyonları içeren web uygulaması Şekil 5'te yer almaktadır.

```python
import streamlit as st
import pandas as pd
from PIL import Image
import subprocess
import os
import base64
import pickle

# Moleküler tanımlayıcı hesaplama fonksyonu
def desc_calc():
    # Tanımlayıcı hesaplamasını gerçekleştirir
    bashCommand = "java -Xms2G -Xmx2G -Djava.awt.headless=true -jar ./PaDEL-Descriptor/PaDEL-Descriptor.jar -removesalt -standardizenitro -fingerprints -descriptortypes ./PaDEL-Descriptor/PubchemFingerprinter.xml -dir ./ -file descriptors_output.csv"
    process = subprocess.Popen(bashCommand.split(), stdout=subprocess.PIPE)
    output, error = process.communicate()
    os.remove('molecule.smi')

# Dosya indirme
def filedownload(df):
    csv = df.to_csv(index=False)
    b64 = base64.b64encode(csv.encode()).decode()  # strings <-> bytes çevirimi
    href = f'<a href="data:file/csv;base64,{b64}" download="p53Tahmin.csv">Tahmini İndirin</a>'
    return href

# Model oluşturma
def build_model(input_data):
    # regression model okuma
    load_model = pickle.load(open('p53_model.pkl', 'rb'))
    # model uygulama
    prediction = load_model.predict(input_data)
```

Şekil 5. Streamlit Web Uygulaması

4. BULGULAR

Hesaplamalı ilaç keşfi, ilaç adayı keşfi ve tasarımında kullanılan hesaplamalı yöntemlerin sağladığı sonuçlar, ilaç keşfi sürecinin etkinliğini artırmada önemli bir rol oynamaktadır. Bu makalede, potansiyel ilaç adaylarının belirlenmesi üzerinde araştırma yapılmıştır ve sonuç olarak p53 proteininin farklı mutasyonları üzerinde hesaplama yapılmıştır. Bu mutasyonlarda hangisinin kimyasal reaksiyonda daha etkili olduğu analiz edilmiştir ve buna uygun makine öğrenmesi modelleri uygulaması geliştirilmiştir. Elde edilen bulgular aşağıda açıklanmıştır. Ayrıca diğer araştırmalarda elde edilen bazı bulgular özet olarak belirtilmiştir.

4.1. Potansiyel ilaç adaylarının belirlenmesi

Hesaplamalı ilaç keşfi yöntemleri, geniş veri kümelerinin analizi ve işlenmesiyle potansiyel ilaç adaylarının belirlenmesinde önemli bir rol oynamaktadır. Bu yöntemler, moleküler yapıların özelliklerini ve aktivitelerini tahmin etmek için kullanılan QSAR (Nicel Yapı-Aktivite İlişkisi) modelleri ve makine öğrenme algoritmaları sayesinde potansiyel ilaç adayları hakkında değerli bilgiler sağlamaktadır. Bu sonuçlar, laboratuvar deneylerine başlamadan önce belirli moleküllerin keşfedilmesi ve seçilmesi sürecinde zaman ve maliyet tasarrufu sağlamaktadır.

4.2. İlaç etkileşimi ve etki mekanizmalarının anlaşılması

Hesaplamalı ilaç keşfi yöntemleri, ilaç moleküllerinin hedef proteinlerle nasıl etkileşime girdiğini ve etki mekanizmalarını anlamak için kullanılmaktadır. Bu yöntemler, moleküler dinamik simülasyonları ve moleküler tarama yöntemlerini içerir. Sonuçlar, ilaç moleküllerinin hedef proteinlere bağlanma şekillerini ve etkileşimlerini ortaya çıkarır. Bu bilgiler, ilaç tasarımı ve Optimizasyonu sürecinde kullanılarak daha etkili ve seçici ilaçlar geliştirilmesine yardımcı olur.

4.3. Yan etkilerin değerlendirilmesi

Hesaplamalı ilaç keşfi yöntemleri, ilaçların yan etki profillerini değerlendirmek için de kullanılmaktadır. Bu yöntemler, ilaç moleküllerinin hedef dışı etkileşimlerini ve toksik etkilerini tahmin etmek için kullanılan modelleri içerir. Sonuçlar, potansiyel yan etkileri olan moleküllerin erken aşamalarda tespit edilmesine ve ilaç tasarımında daha güvenli moleküllerin seçilmesine yardımcı olur.

4.4. İlaç direncinin anlaşılması

Hesaplamalı ilaç keşfi, ilaç direnci ile ilgili sorunları ele almak için de kullanılmaktadır. İlaç direnci, mikroorganizmaların veya kanser hücrelerinin ilaçlara karşı direnç geliştirmesi durumudur. Hesaplamalı yöntemler, ilaç direnci mekanizmalarını anlamak ve bu direncin nasıl aşılabileceğini belirlemek için kullanılmaktadır. Şekil 6'da Biyoaktivite sınıflarının frekans grafik örneğini sunmaktadır. Bu sonuçlar, daha etkili ilaçların geliştirilmesine ve ilaç direnci sorununun üstesinden gelinmesine yardımcı olur. Bunun yanı sıra Şekil 7'de pIC50 değer dağılım grafiği ilaç direncinin anlaşılmasına yardımcı olmaktadır.

4.5. İlaç kombinasyonlarının optimize edilmesi

Hesaplamalı ilaç keşfi, ilaç kombinasyonlarının optimize edilmesinde de önemli bir rol oynamaktadır. Birçok hastalık durumunda, tek bir ilacın etkinliği sınırlı olabilir veya yan etkileri artabilir. Bu nedenle, birden fazla ilacın bir arada kullanılması gerekebilir. Hesaplamalı yöntemler, ilaç kombinasyonlarının etkisini değerlendirmek, etkileşimlerini tahmin etmek ve en uygun kombinasyonları belirlemek için kullanılmaktadır. Bu sonuçlar, daha etkili tedavi yöntemlerinin geliştirilmesine ve hastalıkların tedavisinde daha başarılı sonuçların elde edilmesine yardımcı olur.

Şekil 6. Biyoaktivite Sınıfların Frekans Grafikleri

Sonuç olarak, hesaplamalı ilaç keşfi yöntemleri, ilaç adayı belirlemeden ilaç tasarımına, ilaç etkileşimlerinin analizinden yan etkilerin değerlendirilmesine kadar birçok alanda önemli sonuçlar sunmaktadır. Bu sonuçlar, ilaç keşfi sürecinde hızlı, verimli ve akıllı bir yaklaşımın benimsenmesine olanak tanırken, ilaçların etkinliğini artırmak ve yan etki riskini azaltmak için bilimsel temelli kararlar almayı sağlar. Hesaplamalı ilaç keşfi, gelecekte daha da geliştirilecek ve optimize edilecek olanaklar sunmaktadır ve farmasötik araştırmalara büyük bir potansiyel katkı sağlamaktadır.

Şekil 7. pIC50 Değerlerinin Dağılım Grafiği

Şekil 8'de arayüzü sunulan çalışma, potansiyel ilaç adaylarının belirlenmesine odaklanarak bulgular elde etmiştir. Şekil 9'da proje arayüzüne ait daha kapsamlı gösterim sunulmaktadır. Bulgularımız, potansiyel tedavi seçeneklerini belirlemede büyük potansiyele sahip moleküllerin tanımlanmasıyla ilgili sonuçlardır. Elde edilen sonuçlar şu şekildedir:

Şekil 8. Proje Çalıştırma Gösterimi

Projenin sorunsuz bir şekilde çalışması için, Github **A-Hashemian/final-project** ve **A-Hashemian/final-jupyter** 'den proje klasörünü indirmeniz gerekmektedir. Ardından, proje klasörü içinde bir terminal açarak pip install -r requirements.txt komutunu kullanarak sistem gereksinimlerini indirmeniz önemlidir. Ancak, bazı bilgisayarlarda farklı Python sürümleri bulunması nedeniyle proje her bilgisayarda çalışmayabilir. Bu sorunu çözmek için kendi Python ortamınızı oluşturmanız ve bu ortama Python 3.7.9'u indirmeniz gerekmektedir. Projeyi paylaşabilir bir sunucuda da çalıştırma imkanına sahipsiniz ve projeye aşağıdaki bağlantıdan erişilebilir.
https://a-hashemian-final-project-app-9af9ta.streamlit.app/

Şekil 9. Proje çalıştırma Gösterimi Paylaşabilir Host

5. TARTIŞMA VE SONUÇ

Son on yılda ilaç keşfi ve yapay zeka alanlarında büyük bir gelişim oldu ve bu gelişim hala devam etmektedir. Bununla birlikte, hala ele alınması gereken zorluklar var. Derin öğrenme modellerinin başarısına rağmen, modellerin geliştirilmesi ve değerlendirilmesini en önemli etkeni verilerdir. Aslında doğru veriler seçilmezse model ne kadar gelişmiş olursa olsun doğru sonucu üretemez. Modelleri (tahmini veya üretken) daha kullanışlı hale getirmek için, verilerin yeterli miktarda olması ve yüksek kalitede olması gerekir. Bununla birlikte, önemli bir gerçek şu ki, mevcut kimyasal kütüphaneler büyük miktarda moleküle sahip olsalar da her spesifik tahlil için veri sayısı çok az olabilir. Bazen, kıyaslama veri kümelerinin kalitesi bile, geniş kimyasal alanın dayattığı gerçek dünyadaki ilaç keşfi için temsil gücü açısından sorgulanabilir. İlaç keşfindeki veri kümeleri oldukça dengesiz olabilir. Bu nedenle, modelleri değerlendirirken uygun veri kümelerinin elde edilmesi ve ayrıca veri dengeleme yöntemlerinin yanı sıra uygun değerlendirme ölçütlerinin dikkate alınması gerekir. Ayrıca, ilaç tasarımındaki DMTA döngüsü için her zaman bir ihtiyaç veya belirli hipotezler tarafından yönlendirilmelidir. Mükemmel öngörücü ve üretken modellerle donatılmış olsa bile, bir ilaç adayı tasarlamak için doğru hipotezlerin belirlenmesi gerekir. Başka bir deyişle, ideal bir ilaç adayının özelliklerinin belirlenmesi gereklidir. İlaç tasarımına ilişkin kavrayışlar oluşturmak için, gerçek dünya verilerini (örn. elektronik sağlık kayıtları (EHR) ve ilaç veri tabanları), farklı terapötiklerin etkinliklerini ve yan etkilerini anlamak önem arz etmektedir.
Diğer bir zorluk ta, derin öğrenmenin yorumlama biçiminin insan tarafından idrak edilememesidir. Bu nedenle yorumlanabilir, açıklanabilir makine öğrenmesi modelleri geliştirmek önem arz etmektedir.

Daha spesifik olarak, dört husus önemlidir:

· Sistemin belirli bir cevaba nasıl ulaştığını bilmek olan şeffaflık;
· Model tarafından sağlanan cevabın neden kabul edilebilir olduğunu açıklayan gerekçe;
· İnsan karar vericilere yeni bilgiler sağlayan bilgilendiricilik; ve
· Bir tahminin ne kadar güvenilir olduğunu ölçen belirsizlik tahmini. İdeal bir durum, yapay zekanın bilim insanlarının incelenen süreçle ilgili bilgi ve inançlarını bilmesine izin verebilmesidir.

Bilimsel zorluklara ek olarak, teknik kaygılar devam etmektedir. Göz ardı edilemez bir gerçek şu ki, moleküler grafikler üzerinde öğrenilen en son teknoloji temsili için bile, sabit parmak izleri, moleküler özellik tahmini için GNN den türetilen temsillerden daha iyi performans gösterebilir. Örneğin, değişen hiper-parametre ayarlama, eğitim ve değerlendirme prosedürü bir yana, moleküler özellik tahmini için yapılan çalışmalarda farklı kıyaslama veri kümeleri, farklı bölünmüş katlar ve değerlendirme ölçütleri kullanılmaktadır. Molekül üretimi için Walters, AI tarafından keşfedilen moleküllerin yeniliğini değerlendirmek için kılavuz önermiştir[20].

Aynı şekilde, moleküler özellik tahmini için protokollere de ihtiyaç vardır. Genel olarak, AI'yı ilaç keşfinde uygularken önemli zorlukların yanı sıra birçok umut verici fırsat vardır. Başarılı uygulamalar başlatmak için temel kavramları anlamamız ve amacı, verileri, molekül temsilini, model mimarisini ve öğrenme paradigmasını bir bütün olarak ele almamız gerekmektedir. Bu makalede, yapay zekâ güdümlü ilaç keşfine odaklanan birçok yön ele alınmıştır. Bu yönlerin iyi anlaşılmasıyla, bu alanın önemli ölçüde gelişmesine anlamlı katkılar sağlanacağını öngörüyoruz

Etik Hususlar

Etik kurallara uyum

Çalışmada kullanılan verilerin tamamı açık kaynaklı olarak internette yer almaktadır. İlgili açık kaynaklı veriler literatürde yer almaktadır ve bu çalışmalara temel makaleler referanslara eklenip atıflanmıştır.

Finansman

Çalışma için herhangi bir finansal destek kullanılmamıştır.

Çıkar çatışması

Herhangi bir çıkar çatışması yoktur. Bu çalışma Amin Hashemian'ın lisans 4. Sınıf Tasarım dersinde Doç.Dr. Gıyasettin ÖZCAN danışmanlığında yaptığı çalışma raporundan üretilmiştir.

KAYNAKÇA

[1] Hughes, J. P.; Rees, S., Kalindjian, S. B., Philpott, K. L. 2011. Principles of early drug discovery. Br. J. Pharmacol. 162, 1239–1249.
[2] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. 2019. Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov., 18, 463–477.
[3] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., Blaschke, T. 2018. The rise of deep learning in drug discovery. Drug Discov. Today, 23, 1241–1250.
[4] Mater, A. C., Coote, M. L. 2019. Deep learning in chemistry. J. Chem. Inf. Model, 59, 2545–2559.
[5] Krizhevsky, A., Sutskever, I., Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097–1105.
[6] Mater, A. C., Coote, M. L. 2019. Deep learning in chemistry. J. Chem. Inf. Model, 59, 2545–2559.
[7] Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., et al. 2019. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat. Biotechnol., 37, 1038–1040.
[8] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R.,

French, S., Carfrae, L. A., Bloom-Ackermann, Z., et al. 2020. A deep learning approach to antibiotic discovery. Cell, 180, 688–702.

[9] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. 2019. Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov., 18, 463–477.

[10] LeCun, Y., Bengio, Y., Hinton, G. 2015. Deep learning. Nature, 521, 436–444

[11] Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., G´omez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., Adams, R. P. 2015. Convolutional networks on graphs for learning molecular fingerprints. arXiv preprint arXiv:1509.09292.

[12] Glen, R. C., Bender, A., Arnby, C. H., Carlsson, L., Boyer, S., Smith, J. 2006. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. IDrugs, 9, 199

[13] Goh, G. B., Siegel, C., Vishnu, A. Hodas, N. O., Baker, N. 2017. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert developed QSAR/QSPR models. arXiv preprint arXiv:1706.06689.

[14] Fernandez, M., Ban, F., Woo, G., Hsing, M., Yamazaki, T., LeBlanc, E., Rennie, P. S., Welch, W. J., Cherkasov, A. 2018. Toxic colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images. J. Chem. Inf. Model, 58, 1533–1543.

[15] David, L., Thakkar, A., Mercado, R., Engkvist, O. 2020. Molecular representations in AIdriven drug discovery: a review and practical guide. J. Cheminformatics, 12, 1–22.

[16] Gaulton, A., et al. 2011. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research. 40 (Database issue): D1100-7. doi:10.1093/nar/gkr777. PMC 3245175. PMID 21948594.

[17] Surget, S., Khoury, M.P., Bourdon, J.C. 2013. Uncovering the role of p53 splice variants in human malignancy: a clinical perspective. OncoTargets and Therapy. 7, 57–68. doi:10.2147/OTT.S53876. PMC 3872270. PMID 24379683.

[18] Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J. 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews., 46 (1–3), 3–26. doi:10.1016/S0169-409X(00)00129-0. PMID 11259830.

[19] Lipinski, C.A. 2004. Lead- and drug-like compounds: the rule-offive revolution. Drug Discovery Today: Technologies., 1 (4), 337–341. doi:10.1016/j.ddtec.2004.11.007. PMID 24981612.

[20] Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A., Fisher, J., Jansen, J. M., Duca, J. S., Rush, T. S., et al. 2020. Rethinking drug design in the artificial intelligence era. Nat. Rev. Drug Discov., 19, 353–364.